Introduction to Mira

BG/Q architecture

Early application performance

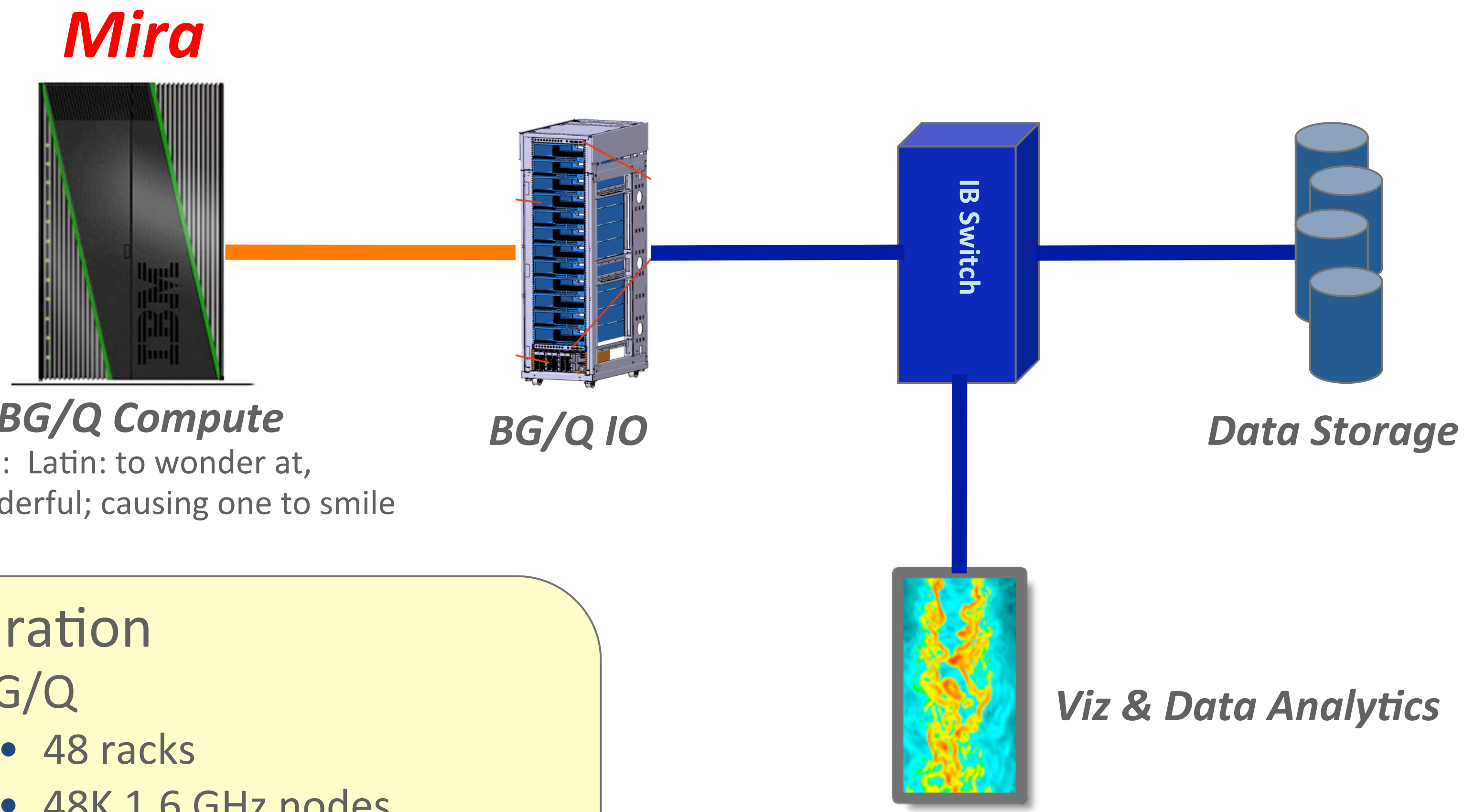
Performance tools, debuggers & libraries

Kalyan Kumaran
ALCF Performance Engineering

New Resources Coming CY2012

- ***Mira* - Blue Gene/Q System**
  - 48K nodes / 768K cores
  - 786 TB of memory
  - Peak flop rate: 10 PF
- **Storage**
  - ~35 PB capacity, 240GB/s bandwidth (GPFS)
  - Disk storage upgrade planned in 2015
    - Double capacity and bandwidth
- **New Visualization Systems**
  - Initial system in 2012
  - Advanced visualization system in 2014
    - State-of-the-art server cluster with latest GPU accelerators
    - Provisioned with the best available parallel analysis and visualization software

ALCF-2: BG/Q System
Mira: A 10PF Computational Science Platform

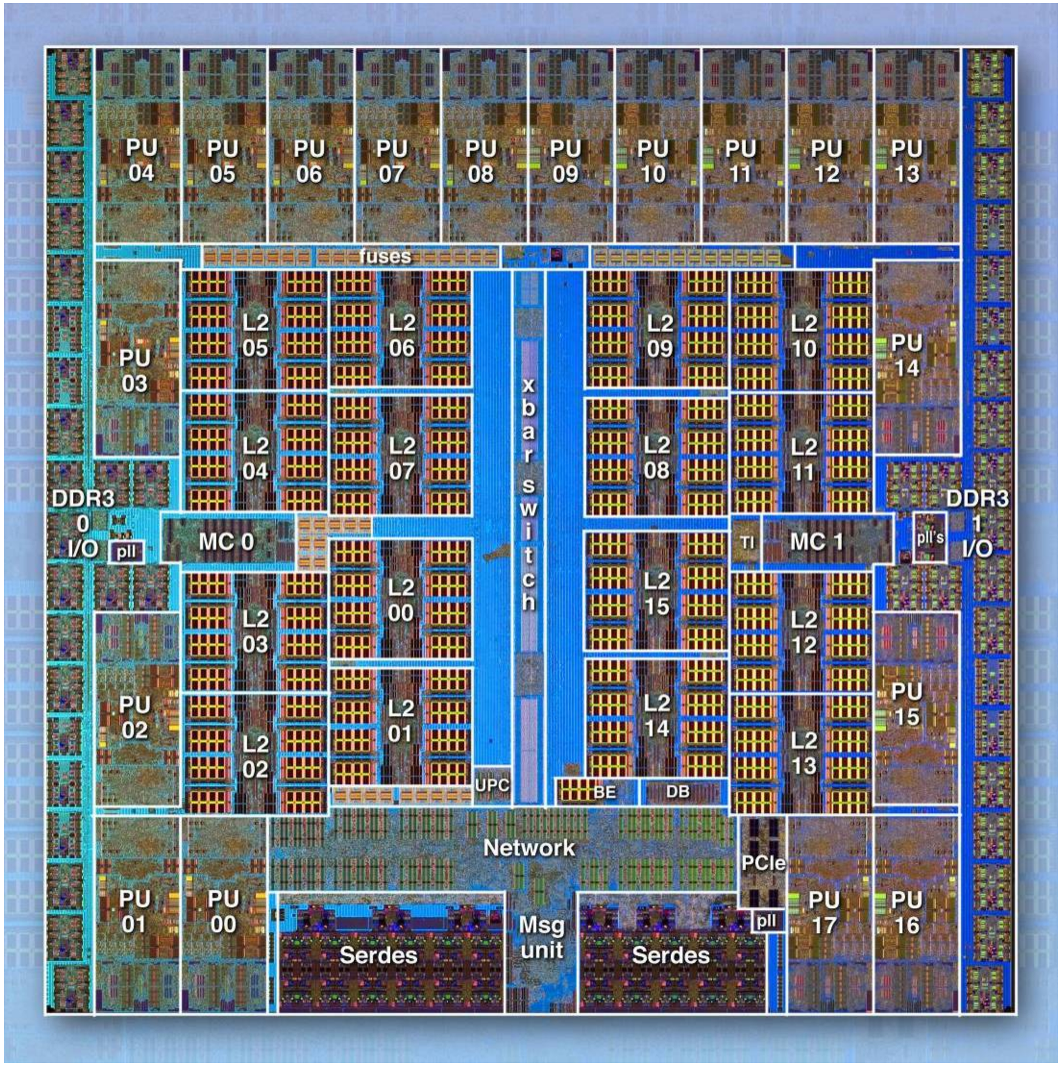


Mira: Latin: to wonder at,
wonderful; causing one to smile

Configuration
- BG/Q
  - 48 racks
  - 48K 1.6 GHz nodes
  - 768K cores & 786TB RAM
  - 384 I/O nodes
- Storage
  - 240 GB/s, 35 PB

BlueGene/Q Compute chip

System-on-a-Chip design : integrates processors, memory and networking logic into a single chip

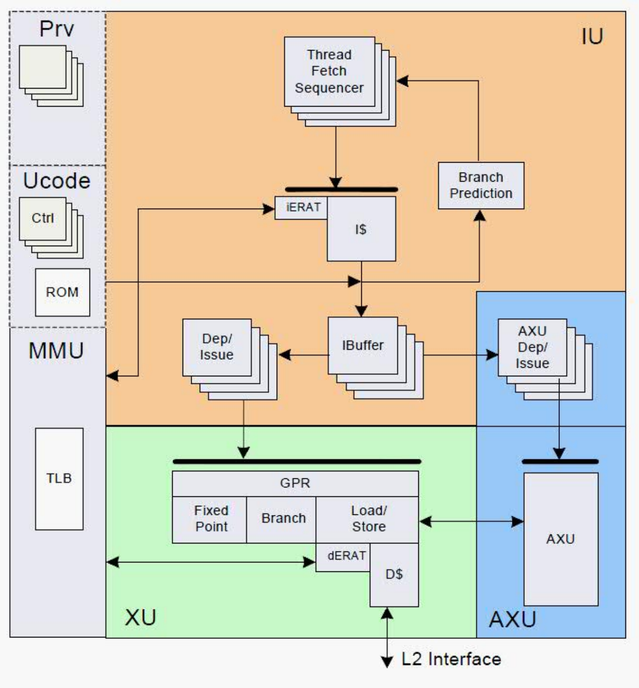


- **360 mm² Cu-45 technology (SOI)**
  - ~ 1.47 B transistors
- **16 user + 1 service processors**
  - plus 1 redundant processor
  - all processors are symmetric
  - each 4-way multi-threaded
  - 64 bits PowerISA™
  - 1.6 GHz
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - each processor has Quad FPU (4-wide double precision, SIMD)
  - peak performance 204.8 GFLOPS@55W
- **Central shared L2 cache: 32 MB**
  - eDRAM
  - multiversioned cache will support transactional memory, speculative execution.
  - supports atomic ops
- **Dual memory controller**
  - 16 GB external DDR3 memory
  - 42.6 GB/s
  - 2 * 16 byte-wide interface (+ECC)
- **Chip-to-chip networking**
  - Router logic integrated into BQC chip.
- **External IO**
  - PCIe Gen2 interface

BG/Q Processor Unit

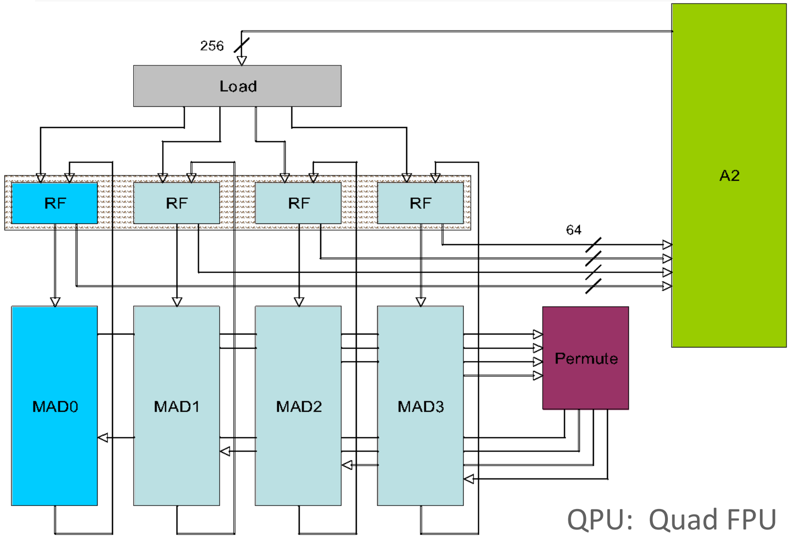
A2 processor core

- Mostly same design as in PowerEN™ chip
- Implements 64-bit PowerISA™
- Optimized for aggregate throughput:
  - 4-way simultaneously multi-threaded (SMT)
  - 2-way concurrent issue 1 XU (br/int/l/s) + 1 FPU
  - in-order dispatch, execution, completion
- L1 I/D cache = 16kB/16kB
- 32x4x64-bit GPR
- Dynamic branch prediction
- 1.6 GHz @ 0.8V



Quad FPU

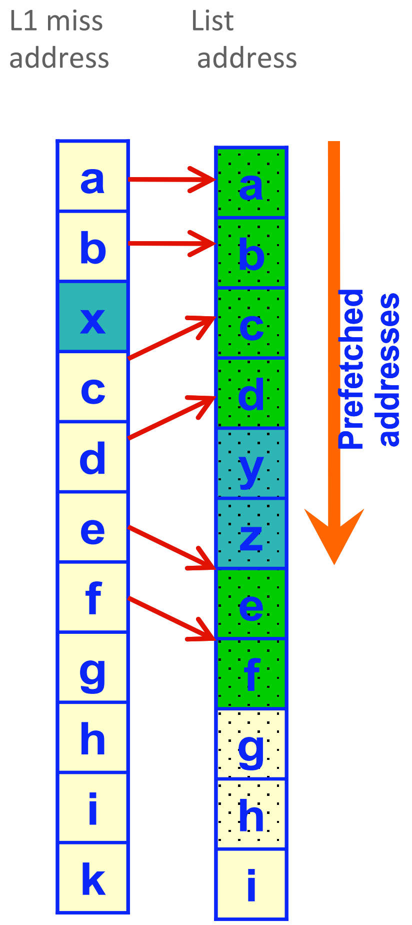
- 4 double precision pipelines, usable as:
  - scalar FPU
  - 4-wide FPU SIMD
  - 2-wide complex arithmetic SIMD
- Instruction extensions to PowerISA
- 6 stage pipeline
- 2W4R register file (2 * 2W2R) per pipe
- 8 concurrent floating point ops (FMA) + load + store
- Permute instructions to reorganize vector data
  - supports a multitude of data alignments



QPU: Quad FPU

BlueGene/Q PUnit - ct.

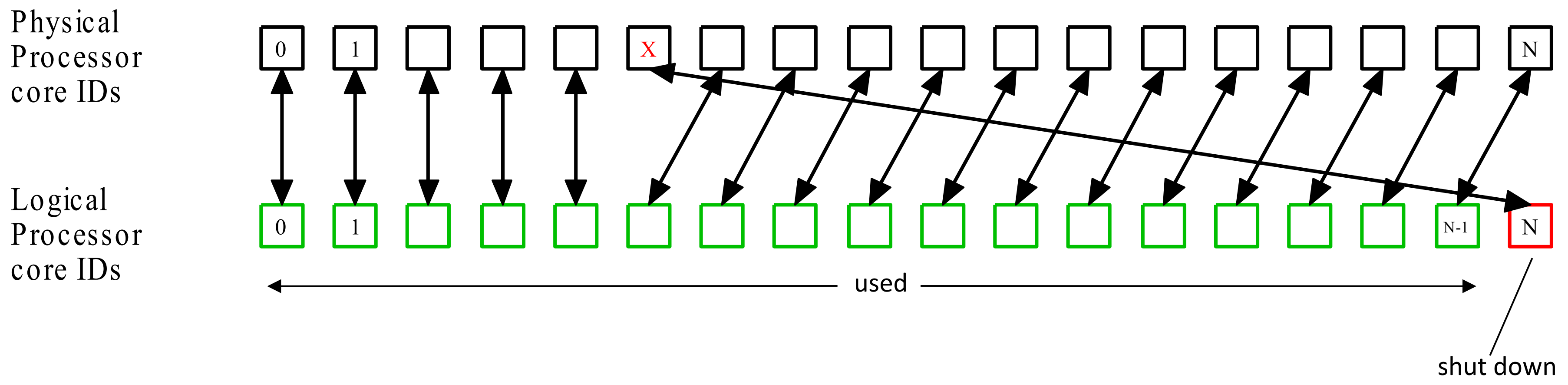
- **L1 prefetcher**
  - Normal mode: Stream Prefetching
    - in response to observed memory traffic, adaptively balances resources to prefetch L2 cache lines (@ 128 B wide)
    - from 16 streams x 2 deep through 4 streams x 8 deep
  - Additional: 4 List-based Prefetching engines:
    - One per thread
    - Activated by program directives, e.g. bracketing complex set of loops
    - Used for repeated memory reference patterns in arbitrarily long code segments
    - Record pattern on first iteration of loop; playback for subsequent iterations
    - On subsequent passes, list is adaptively refined for missing or extra cache misses (async events)



List-based "perfect" prefetching has tolerance for missing or extra cache misses

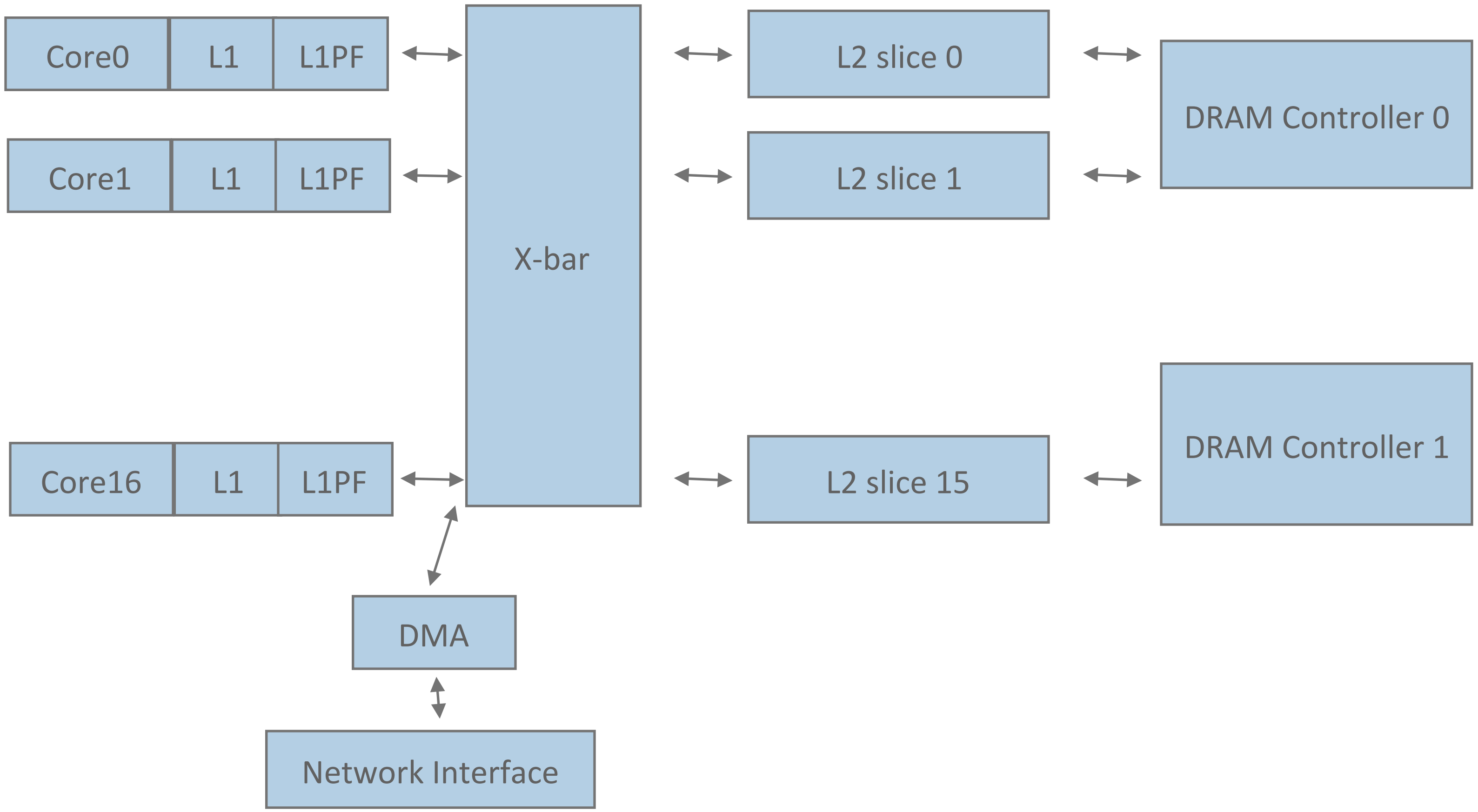
- **Wake-up unit**
  - Will allow SMT threads to be suspended, while waiting for an event
  - Lighter weight than wake-up-on-interrupt -- no context switching
  - Improves power efficiency and resource utilization

Physical-to-Logical mapping of PUnits in presence of a fail



- Inspired by array redundancy
- PUnit N+1 redundancy scheme substantially increases yield of large chip
- Redundancy can be invoked at any manufacturing test stage
  - wafer, module, card, system
- Redundancy info travels with physical part -- stored on chip (eFuse) / on card (EEPROM)
  - at power-on, info transmitted to PUnits, memory system, etc.
- Single part number flow
- Transparent to user software: user sees N consecutive good processor cores.

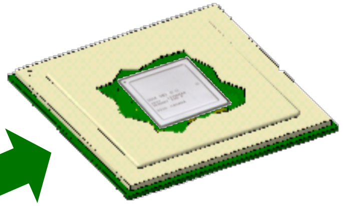
BG/Q Memory Structure



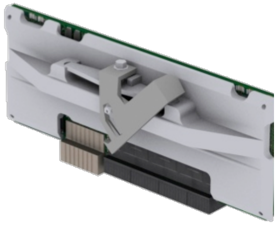
Blue Gene/Q

1. Chip:
16+2 □P
cores

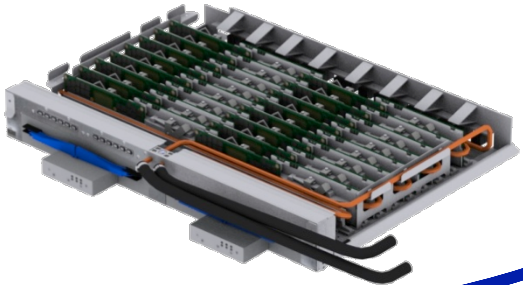
2. Single Chip Module



3. Compute card:
One chip module,
16 GB DDR3 Memory,
Heat Spreader for H_2O Cooling



4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips; 5D Torus

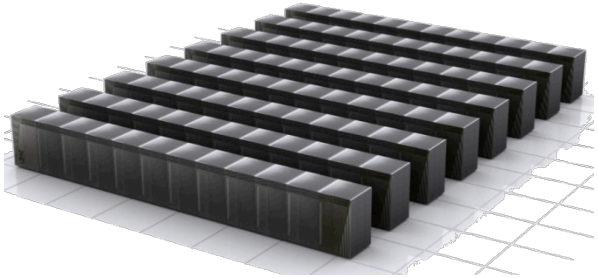


5a. Midplane:
16 Node Cards

5b. IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots
3D I/O torus

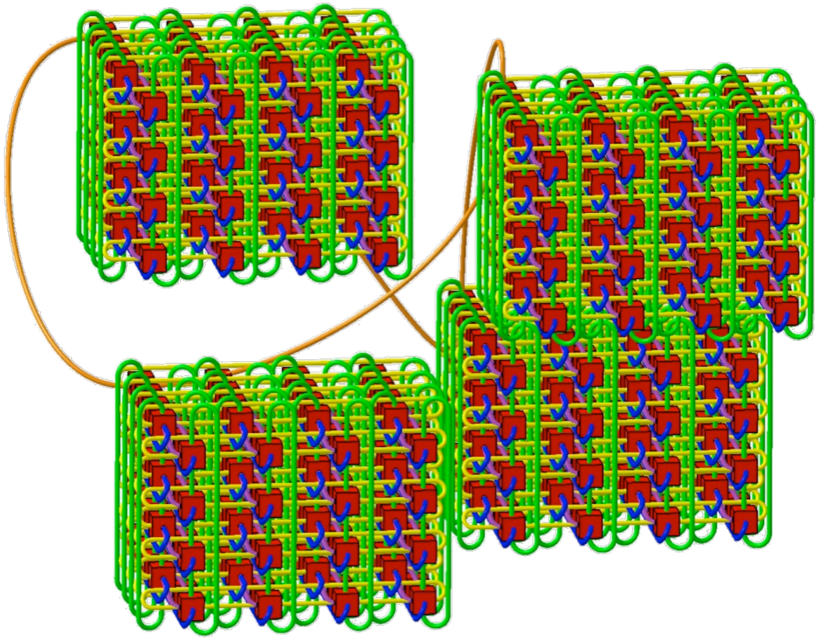
6. Rack: 2 Midplanes

7. System:
96 racks, 20PF/s



- **Sustained single node perf: 10x P, 20x L**
- **MF/Watt: (6x) P, (10x) L (~2GF/W, Green 500 criteria)**
- **Software and hardware support for programming models for exploitation of node hardware concurrency**

Inter-Processor Communication

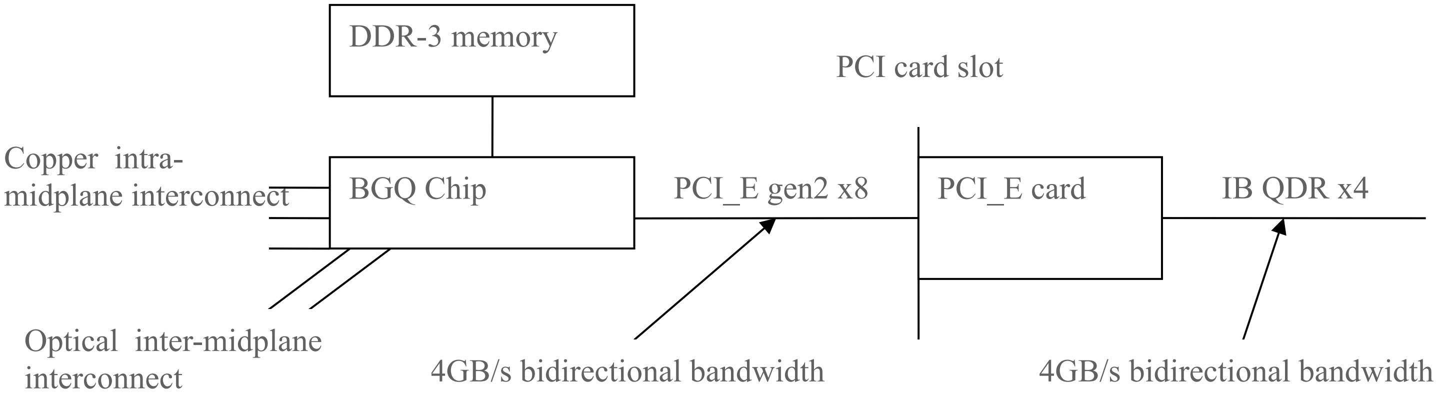


Network Performance

- **All-to-all: 97% of peak**
- **Bisection: > 93% of peak**
- **Nearest-neighbor: 98% of peak**
- **Collective: FP reductions at 94.6% of peak**

- **Integrated 5D torus**
  - Virtual Cut-Through routing
  - Hardware assists for collective & barrier functions
  - FP addition support in network
  - RDMA
    - Integrated on-chip Message Unit
- **2 GB/s raw bandwidth on all 10 links**
  - each direction -- i.e. 4 GB/s bidi
  - 1.8 GB/s user bandwidth
    - protocol overhead
- **5D nearest neighbor exchange measured at 1.76 GB/s per link (98% efficiency)**
- **Hardware latency**
  - Nearest: 80ns
  - Farthest: 3us
    (96-rack 20PF system, 31 hops)
- **Additional 11th link for communication to IO nodes**
  - BQC chips in separate enclosure
  - IO nodes run Linux, mount file system
  - IO nodes drive PCIe Gen2 x8 (4+4 GB/s)
    ↔ IB/10G Ethernet ↔ file system & world

Blue Gene/Q I/O node

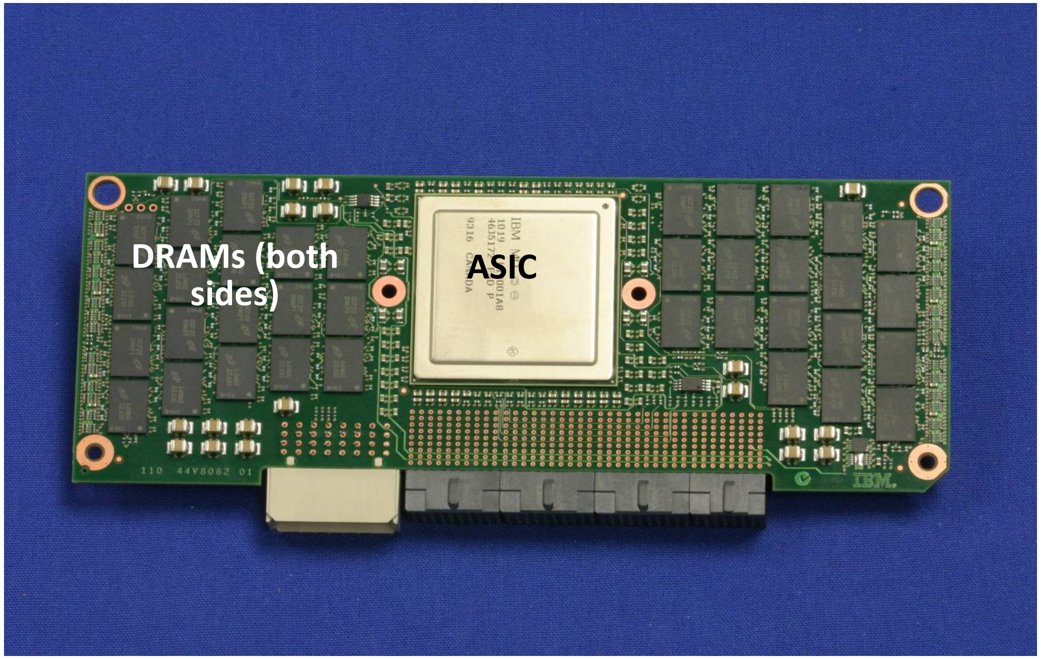

DDR-3 memory
PCI card slot
Copper intra-midplane interconnect
BGQ Chip
PCI_E gen2 x8
PCI_E card
IB QDR x4
Optical inter-midplane interconnect
4GB/s bidirectional bandwidth
4GB/s bidirectional bandwidth

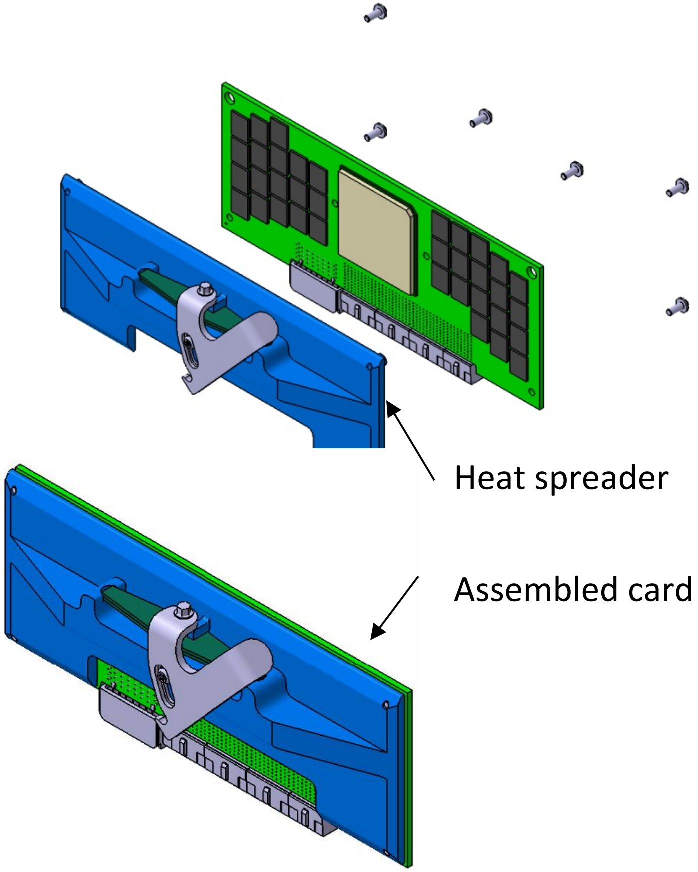
Alternatives:
-- PCI_E to IB QDR x4 (shown)
-- PCI_E to (dual) 10 Gb ethernet card (log in nodes)
-- PCI_E to single 10GbE + IB QDR
-- PCI_E to SATA for direct disk attach

BG I/O Max Bandwidth

	BG/L	BG/P	BG/Q
Type	1GbE	10GbE	PCI-e
BW/node	1Gb/s x2 250MB/s	10Gb/sx2 2.5GB/s	4GB/sx2
# of I/O nodes	128	64	8-128
BW/rack in BW/rack out	16GB/s 16GB/s	80GB/s 80GB/s	512GB/s@128 512GB/s@128
I/O byte/flop	0.0056	0.011	0.0048

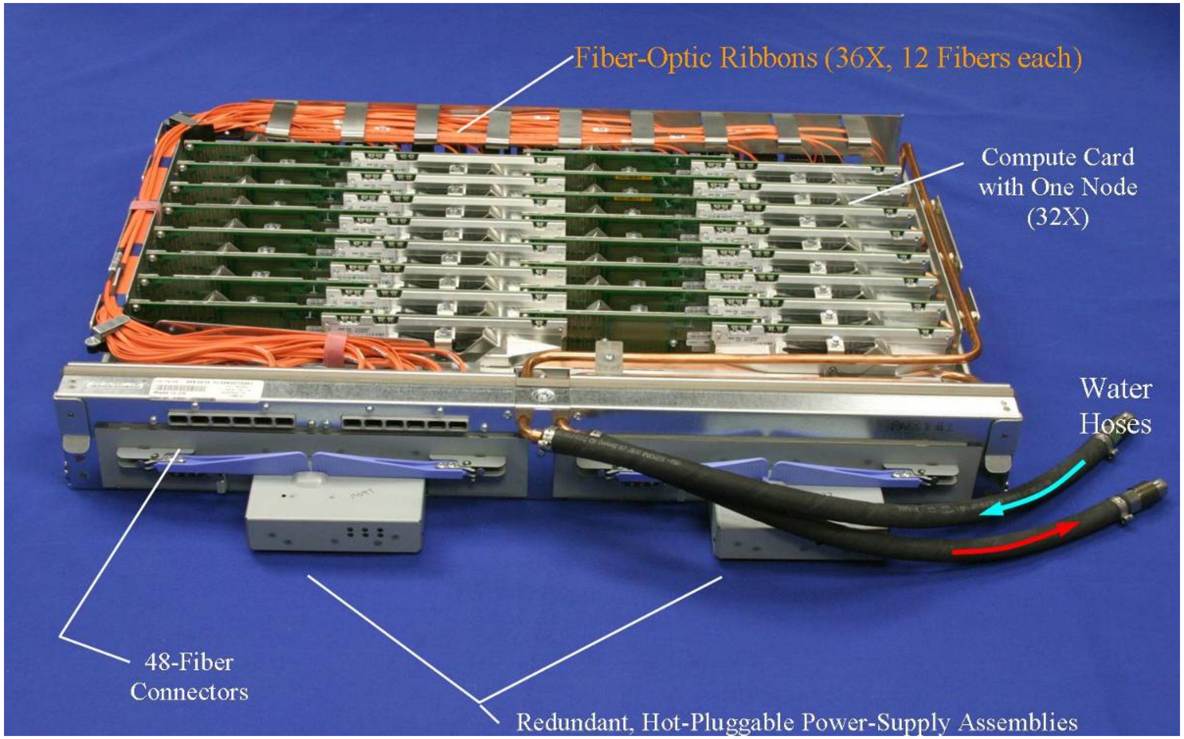
Blue Gene/Q Compute Card Assembly





- Basic field replaceable unit of a Blue Gene/Q system
- Compute Card has 1 BQC chip + 72 SDRAMs (16GB DDR3)
- Two heat sink options: Water-cooled → "Compute Node" / air-cooled → "IO Node"
- Connectors carry power supplies, JTAG etc, and 176 Torus signals (4 and 5 Gbps)

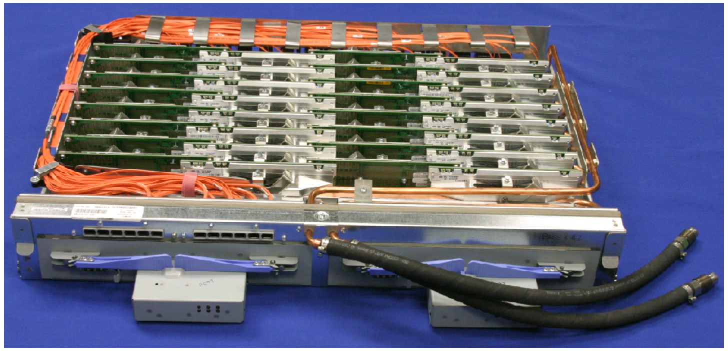
Blue Gene/Q Node Card Assembly


Fiber-Optic Ribbons (36X, 12 Fibers each)
Compute Card with One Node (32X)
Water Hoses
48-Fiber Connectors
Redundant, Hot-Pluggable Power-Supply Assemblies

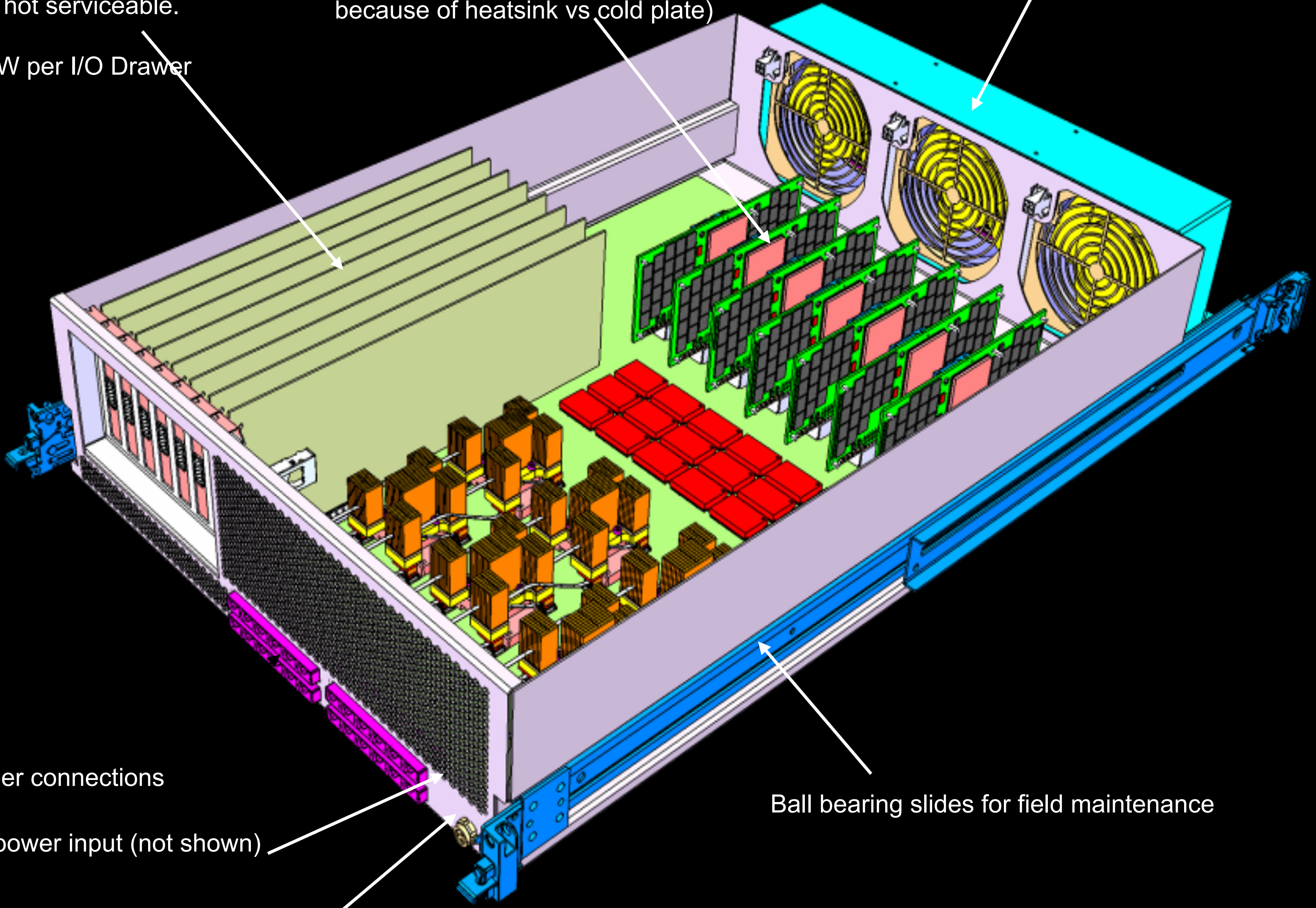
- **Power efficient processor chips allow dense packaging**
- **High bandwidth / low latency electrical interconnect on-board**
- **18+18 (Tx+Rx) 12-channel optical fibers @10Gb/s**
  - Recombined into 8*48-channel fibers for rack-to-rack (Torus) and 4*12 for Compute-to-IO interconnect
- **Compute Node Card assembly is water-cooled (18-25ºC – above dew point)**
- **Redundant power supplies with distributed back-end ~ 2.5 kW**

Packaging and Cooling

Water	18C to 25C
Flow	20 gpm to 30 gpm
Height	2095 mm (82.5 inches)
Width	1219 mm (48 inches)
Depth	1321 mm (52 inches)
Weight	2000 kg (4400 lbs) *(including water)*
	I/O enclosure with 4 drawers 210 kg (480 lbs)



- Water cooled node board
- 32 compute cards, 8 link ASICs drive 4D links using 10Gb/s optical transceivers
- Hot pluggable front-end power supplies



Full height, 25W PCI cards,
NOT hot serviceable.

~1 KW per I/O Drawer

8 compute cards
(different PN than in compute rack
because of heatsink vs cold plate)

Axial fans – same as BGP

Fiber connections

48V power input (not shown)

Clock input

Ball bearing slides for field maintenance

Picture by Shawn Hall

Overview of BG/Q: Another step forward

Design Parameters	BG/P	BG/Q	Improvement
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
Nodes / Rack	1,024	1,024	--
RAM / core (GB)	0.5	1	2x
Flops / Node (GF)	13.6	204.8	15x
Mem. BW/Node (GB/sec)	13.6	42.6	3x
Latency (MPI zero-length, nearest-neighbor node)	2.6 µs	2.2 µs	~15% less
Bisection BW (32 racks)	1.39TB/s	13.1TB/s	9.42x
Network Interconnect	3D torus	5D torus	Smaller diameter
Concurrency / Rack	4,096	65,536	16x
GFlops/Watt	0.77	2.10	3x

BG/Q A2 Core – Quick Overview for the Programmer

- Full PowerPC compliant 64-bit CPU (BG/P PowerPC 450d was 32-bit)
- 1.6GHz, in-order execution, 4 hardware threads/core, 16 cores/node, 16GB/node
- At most one instruction can be completed per cycle per thread
- At most 2 instructions can be completed per cycle per core:
  - one instruction must be integer/load/store (XU)
  - one instruction must be floating point (AXU)
- 4-wide SIMD floating point unit with complete set of parallel instructions
  - 4 FMA's @ 1.6GHz = 12.8 Gflops/core
- Cache:
  - 16 KB L1 data cache, 64 byte lines, shared between 4 hardware threads
  - L1 Prefetch buffer, 32 lines, 128 bytes each
  - 32 MB shared L2 cache

Notes on Mira Science Applications

- Applications cannot be manually tuned; only compiler optimizations are allowed.
- 3 of the applications are threaded – i.e., use both OpenMP and MPI (GTC, GFMC, GFDL).
- The remainder are 100% MPI applications (DNS3D, FLASH, GPAW, LS3DF, MILC, NAMD & NEK 5000).
- For 100% MPI applications, we tested multiple MPI ranks per core (max of 4 ranks per core).
- For MPI + OpenMP applications, we tested 1 MPI rank per core and multiple OpenMP threads per core (max of 4 threads per core)

Comments on using all hardware threads

- Speed up with hardware threads will be limited if the issue rate is already high with 1 thread/core (NEK is an example).
- Speed-up with hardware threads will be limited if the problem is already near the scaling limit at 1 thread/core. Using all threads will require 4x more threads.
- Speed-up can be limited if there is contention for L1-D and L1P resources.
- In some cases using OpenMP or Pthreads instead of MPI might reduce L1 contention.

BG/Q Performance Tools

- **Early efforts were initiated to bring widely used performance tools to the BG/Q**
- **A variety of tools providers are currently working with IBM and Argonne to port and test tools on the Q**
- **BG/Q provides a hardware &software environment that supports many standard performance tools:**
  - Software:
    - Environment similar to 64 bit PowerPC Linux
      - provides standard GNU binutils
    - New performance counter API bgpm
  - Performance Counter Hardware:
    - BG/Q provides 424 64-bit counters in a central node counting unit
    - Counter for all cores, prefetchers, L2 cache, memory, network, message unit, PCIe, DevBus, and CNK events
    - Provides support for hardware threads and counters can be controlled at the core level
    - Countable events include: instruction counts, flop counts, cache events, and many more

BG/Q Tools

Tool Name	Source	Provides	Q Status
gprof	GNU/IBM	Timing (sample)	In development
TAU	Unv. Oregon	Timing (inst), MPI	Development pending
Rice HPCToolkit	Rice Unv.	Timing (sample), HPC (sample)	In development & testing
IBM HPCT	IBM	MPI, HPC	In development
mpiP	LLNL	MPI	In development & testing
PAPI	UTK	HPC API	In development & testing
Darshan	ANL	IO	In development & testing
Open|Speedshop	Krell	Timing (sample), HCP, MPI, IO	In development
Scalasca	Juelich	Timing (inst), MPI	In development & testing
FPMPI2	UIUC	MPI	Development planned
DynInst	UMD/Wisc/IBM	Binary rewriter	In development
ValGrind	ValGrind/IBM	Memory & Thread Error Check	Development planned

Parallel Debuggers

- **IBM CDTI (Code Development and Tools Interface)**
  - Collaboration of IBM/LLNL/ANL resulted in update v1.7 (August 2011)
  - Refined interface for multiple tool support, breakpoint handling, stepping, and signal handling
- **Rogue Wave TotalView**
  - Ported to BG/Q (Q32 at IBM) with basic functionality in August 2011
  - Pre-release testing by LLNL December 2011
  - Status
    - Tested working: basic ops (step, breakpoint, stack), QPX instructions, fast conditional breakpoints, job control for C/C++/Fortran with MPI/OMP/threads.
    - Still testing: **scalability**, fast conditional watchpoints, debugging in TM/SE
- **Allinea DDT**
  - Preparation via ANL scalability research contract on BG/P to address I/O node bottlenecks
    - Multiplexed debug daemons – complete and tested (Nov 2011)
    - Multiplexed gdbserver processes – complete and tested for single threading (Dec 2011)
    - Still testing: multiplexed gdbserver with multiple threads/process.
  - Status
    - Expected BG/P Beta release Jan 2012.
    - BG/Q port to begin on ANL T&D Feb 2012 as part of Early Science project (ESP).

Libraries

- ESSL available through IBM
- PETSc is being optimized as part of BG/Q Tools ESP project
- Will port and tune 3rd party libraries (FFTW, BLAS, LAPACK, ScaLAPACK, ParMetis, P3DFFT, …) using compiler optimizations
- Collecting actual library usage data; libraries will be stamped with a detectable string id.
- Collaborating with Robert van de Geijn's group on rewriting Goto-BLAS so that it can be easily ported and tuned to new architectures like BG/Q (BLIS)
- Exploring an optimized FFT library with Spiral Gen

Math Libraries in /soft/libraries/alcf

- Maintained in-house, frequently updated
- GCC and XL built versions of each library
- BLAS
- LAPACK 3.3.1
- ScaLAPACK 2.0.2
- FFTW 2.1.5
- FFTW 3.3.1
- PARPACK

Math Library Future Plans

- LAPACK 3.4.1 port (with LAPACKE)
- CBLAS
- METIS/ParMETIS
- Goto-BLAS ported and tuned on BG/Q
- New kernel infrastructure codenamed “BLIS”, designed in collaboration with Univ. of Texas
- Right now ESSL GEMM routines are extracted into the ALCF BLAS library
- Tune FFTW 3.x, time permitting

Libraries in /soft/libraries/unsupported

- Not actively maintained by ALCF (at least for now)
- Provided as a convenience
- Boost 1.49.0
- HDF5 1.8.8
- NETCDF 4.1.3
- P3DFFT 2.4 (patched)
- Tcl 8.4.14
- zlib 1.2.6